A Formal Analysis of Bias in User-Agnostic Ranking Systems

Guilherme Ramos
Instituto de Telecomunicações, and Instituto Superior Técnico, ULisboa
1049-001 Lisboa, Portugal
guilherme.ramos@tecnico.ulisboa.pt

Ludovico Boratto
University of Cagliari
Cagliari, Italy
ludovico.boratto@acm.org

Mirko Marras
University of Cagliari
Cagliari, Italy
mirko.marras@acm.org

ABSTRACT

Recent advances in ranking systems have shifted towards user-agnostic approaches to address ethical concerns regarding user profiling. While designed to be non-discriminatory, the fairness properties of such systems require empirical validation. Using the User-Agnostic Ranking System (UARS) as a case study, we analyze its bias across sex and age groups through a novel bias metric that quantifies disparities in rating preservation. Our empirical evaluation spans three diverse datasets—MovieLens-100k, MovieLens-1M, and BookCrossing—highlighting contrasting patterns: UARS exhibits low sex-based bias (0.09 to 0.16) but significant and increasing age-based bias for extreme age groups (up to 0.98). These findings reveal that statistical filtering mechanisms, while avoiding explicit user profiling, may still propagate demographic disparities, underscoring the need for refined design principles in truly equitable ranking systems.

CCS CONCEPTS

• **Information systems** → **Learning to rank**; • **Applied computing** → *Law, social and behavioral sciences.*

KEYWORDS

Ranking System, user agnostic, fairness, non-discrimination.

1 INTRODUCTION

The evolution of contemporary society into an information-driven economy has fundamentally transformed how online platforms influence user decisions and item success. Ranking systems play a crucial role in this landscape, with their impact often exceeding traditional marketing strategies [1, 2, 4, 9]. Recent regulatory frameworks, particularly the EU AI Act [13], have highlighted the ethical concerns surrounding systems that evaluate individuals based on behavior or predicted attributes [22], leading to a critical examination of traditional reputation-based ranking approaches [3, 10, 12].

The User-Agnostic Ranking System (UARS) was recently proposed as a solution that aims to achieve fair item ranking through statistical filtering rather than user reputation scores [18]. By design, UARS processes ratings based purely on their statistical properties without considering user attributes, theoretically offering a path to non-discrimination. While this approach aligns with emerging regulations and maintains system effectiveness [7, 19, 20], its fairness properties require careful examination. Traditional countermeasures involving user reputation scoring [10, 12] have sparked ethical concerns around user discrimination [16, 17] and misuse of personal data [22].

While UARS shows promising properties in terms of manipulation resistance [18], its claims of inherent fairness require rigorous empirical validation. The relationship between statistical filtering mechanisms and potential bias remains unexplored, particularly concerning sensitive attributes such as sex, age, or occupation [14, 16]. Understanding these aspects is crucial as ranking systems increasingly influence social and economic outcomes [8, 11, 15, 21, 23].

In this paper, we provide a systematic analysis of potential bias in UARS across multiple sensitive attributes. We demonstrate through theoretical analysis and empirical validation on three real-world datasets (MovieLens-100k, MovieLens-1M, and Amazon Musical Instruments) that UARS's statistical filtering mechanism inherently prevents systematic bias from propagating to final rankings. Our key findings show that:

- UARS demonstrates relatively low sex-based bias across datasets, with bias scores ranging from 0.09 to 0.16.
- Age-based bias shows systematic patterns, with higher bias scores for extreme age groups (up to 0.98 for users over 55).
- The effectiveness of statistical filtering varies significantly across demographic segments, suggesting that user-agnostic design principles alone may not guarantee fairness.

These results contribute to a deeper understanding of how algorithmic design choices affect fairness properties in ranking systems, revealing both the potential and limitations of statistical filtering approaches. Our analysis extends the growing body of work on fair and robust ranking systems [13, 16, 17, 22]. Furthermore, our analysis extends the growing body of work on fair and robust ranking systems [14, 16, 19] while highlighting the complexity of achieving true non-discrimination in practice [5, 7].

SIGIR '25, July 13–18, 2025, Padua, Italy

ACM ISBN –.
https://doi.org/--

1.1 Notation

Next, in Table 1, we summarize the adopted notation in this work.

Table 1: Notation used throughout the paper.

Symbol	Description
U	Set of all users
I	Set of all items
U_i	Set of users who rated item i
R_i	Set of valid ratings for item i (those that have not been filtered out by the statistical filter)
$\|U_i\|$	Total number of users who rated item i
$\|R_i\|$	Number of ratings that remain after filtering
μ_i	Mean of valid ratings for item i: $\mu_i = \frac{\sum_{r \in R_i} r}{\|R_i\|}$
σ_i	Standard deviation of valid ratings for item i: $\sigma_i = \sqrt{\frac{\sum_{r \in R_i} (r-\mu_i)^2}{\|R_i\|-1}}$
R_{ui}	Rating given by user u to item i
$\neg g$	denotes the complement of users in demographic group g

2 BACKGROUND: STATISTICAL FILTERING IN UARS

To rigorously examine whether UARS truly achieves its promise of non-discrimination, we need to analyze how its statistical filtering mechanism affects different demographic groups. Our analysis particularly focuses on two key dimensions: sex, where initial evidence suggests relatively low bias, and age groups, where theoretical concerns about varying rating patterns might emerge. The key question we address is: Does statistical filtering treat ratings from different demographic groups fairly, even when these groups might exhibit systematically different rating behaviors?

2.1 Rating Processing Mechanism

UARS implements an iterative statistical filtering process centered on outlier detection and removal. For each item i, the system maintains a dynamic set of valid ratings R_i, derived from the initial set of all ratings $\{R_{ui} : u \in U_i\}$. The process computes the mean $\mu_i = \frac{\sum_{r \in R_i} r}{|R_i|}$ and standard deviation $\sigma_i = \sqrt{\frac{\sum_{r \in R_i} (r-\mu_i)^2}{|R_i|-1}}$ of the current valid ratings. It then applies the statistical filtering criterion $(R_{ui} - \mu_i)^2 \leq \sigma_i$ to determine which ratings remain in the valid set. This computation repeats until the set of valid ratings stabilizes, at which point the final mean becomes the item's score. The schematics of UARS is depicted in Figure 1.

2.2 Core Properties and Implications

The design of UARS embeds statistical objectivity and user anonymity as fundamental principles. The system evaluates ratings based purely on their distributional properties, deliberately ignoring user identities and attributes during processing. This convergent filtering approach ensures stable rating sets while maintaining identity-blind rankings that depend only on rating patterns, not user characteristics.

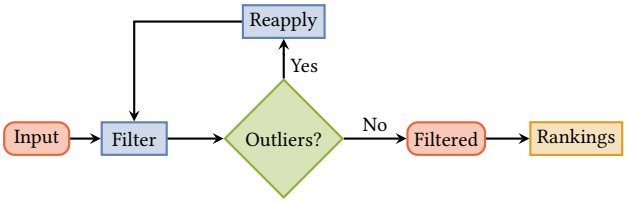

Figure 1: Diagram representing the filtering steps of the UARS.

While these properties suggest inherent unbiased behavior, the impact of statistical filtering on fairness requires deeper examination, particularly when different demographic groups exhibit distinct rating patterns. The relationship between statistical validity and demographic fairness remains an open question that our analysis aims to address. The following sections present our methodology for investigating whether UARS's identity-blind approach effectively prevents discriminatory outcomes across diverse user populations.

3 METHODOLOGY

To rigorously examine whether UARS truly achieves its promise of non-discrimination, we need to look beyond its theoretical properties and understand how it behaves with real-world data and diverse user populations. The key question we address is: Does statistical filtering treat ratings from different demographic groups fairly, or might it inadvertently favor certain groups over others?

3.1 Quantifying Bias in Statistical Filtering

At its core, UARS makes decisions about which ratings to keep and which to filter out based purely on statistical properties. However, different demographic groups might have systematically different rating patterns. For instance, older users might tend to provide more extreme ratings, or younger users might cluster their ratings differently. While UARS's statistical filtering is blind to user attributes, such systematic differences in rating behavior could lead to disparate treatment.

We propose a straightforward but powerful metric that captures potential disparities in how UARS treats different groups. For an item i and a demographic group g with at least one user (such as female users or users in a particular age range), we first calculate the proportion of ratings that survive the filtering process:

$$P_g(i) = \frac{|R_g(i)|}{|U_g(i)|}. \tag{1}$$

Here, $R_g(i)$ represents the ratings from group g that UARS deemed valid (not outliers), while $U_g(i)$ represents all ratings from that group. This ratio tells us what fraction of a group's voices are preserved in the final ranking.

To detect potential bias, we compare this proportion between a group and its complement:

$$\text{Bias}_g(i) = \begin{cases} 0 & \text{if } |U_g(i)| = 0 \text{ or } |U_{\neg g}(i)| = 0 \\ \left|P_g(i) - P_{\neg g}(i)\right| & \text{otherwise,} \end{cases} \quad (2)$$

where $P_g(i)$ and $P_{\neg g}(i)$ are defined by Equation (1). This definition ensures that when a demographic group has no ratings for an item, we consider there to be no measurable bias, since we cannot meaningfully compare rating preservation ratios. Moreover, a bias score of zero represents perfect equity - both groups have the same proportion of their ratings preserved. Higher scores indicate potential systematic differences in how UARS treats different groups.

To interpret the bias metric, we consider a bias score of 0 as perfectly equitable, meaning no systematic differences exist between groups. However, determining what constitutes an "acceptable" bias score depends on the application context and societal benchmarks. For instance, regulatory frameworks such as the EU AI Act emphasize minimizing algorithmic bias but do not specify numerical thresholds. Drawing inspiration from prior work in algorithmic fairness, we consider bias scores below 0.1 to indicate very low bias, scores between 0.1 and 0.3 as low bias, scores between 0.3 and 0.5 as moderate bias, and above that as potentially concerning. These thresholds provide a starting point for evaluation but warrant further exploration in domain-specific contexts. In our analysis, we pay particular attention to:

- Systematic patterns in bias scores across demographic groups.
- Relative magnitudes of bias between different attributes (e.g., sex vs. age).
- Consistency of bias patterns across different datasets.

A bias score of zero represents perfect equity - both groups have the same proportion of their ratings preserved. Higher scores indicate potential systematic differences in how UARS treats different groups.

3.2 From Individual Items to System-wide Analysis

While examining bias for individual items provides valuable insights, we need to understand the system's behavior as a whole. We analyze several aggregate measures:

- Mean bias ($\overline{\text{Bias}_g}$): Captures the typical level of disparity across all items.
- Bias distribution: Reveals whether disparities are consistent or concentrated in certain items.

3.3 Empirical Validation

To ensure our findings are robust and generalizable, we conduct our analysis across three diverse datasets that represent different scales and domains:

Our dataset selection aims to validate bias patterns across different contexts:

- **MovieLens-100k** [6]: A focused dataset with rich demographic information, including sex, age, and occupation attributes, allowing detailed analysis of multiple bias dimensions. It contains 100,000 ratings of $|I| = 1{,}682$ movies made by $|U| = 943$ users.
- **MovieLens-1M** [6]: A larger-scale validation that maintains the same demographic attributes as ML-100k. It contains 1,000,209 ratings of $|I| = 3{,}952$ movies made by $|U| = 6{,}040$ users.
- **Book-Crossing** [24]: A book rating platform dataset containing age and location attributes, representing a distinct domain from movie ratings. While location information is available, we excluded it from our analysis due to data sparsity (hundreds of different countries) and quality issues (numerous unintelligible or inconsistent location entries), focusing instead on the more reliable age attribute. It contains 745,161 ratings, $|U| = 53{,}408$ users, and $|I| = 263{,}956$ items.

In Figure 2, we depict the distribution of users per atribute in each dataset. This diverse selection helps distinguish between dataset-specific effects and inherent properties of the UARS mechanism. Particularly, the inclusion of both movie and book ratings allows us to examine whether the observed bias patterns persist across different types of content and user behaviors. While MovieLens datasets provide comprehensive demographic coverage, BookCrossing offers complementary insights focused specifically on age-based bias in a different cultural context.

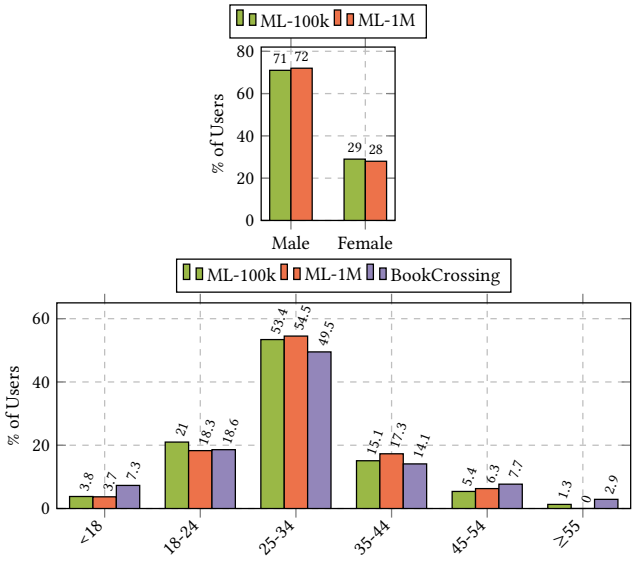

Figure 2: Attributes distribution. Above: sex distribution comparison between MovieLens datasets. Despite the different scales of the datasets, they show similar gender distributions. Below: age distribution comparison between all datasets. Similar age distributions.

For each dataset, we examine multiple demographic dimensions, namely sex and age groups. This multi-faceted analysis helps us understand whether UARS's fairness properties hold consistently across different types of demographic attributes and user populations.

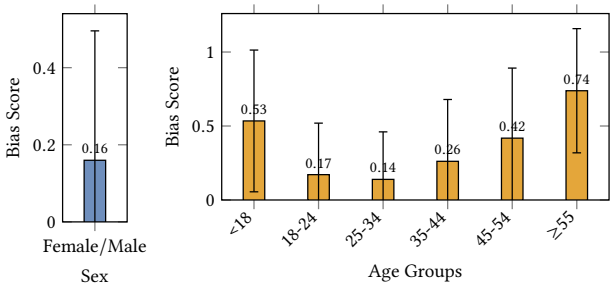

Figure 3: Average bias scores across demographic groups – ML-100k. Left: Sex-based bias score for female/male users. Right: Age-based bias scores across different age groups. Error bars represent standard deviation.

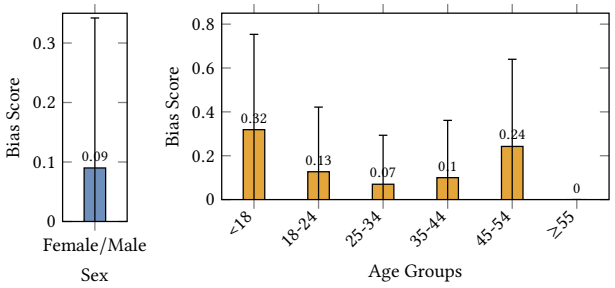

Figure 4: Average bias scores across demographic groups – ML-1M. Left: Sex-based bias score for female/male users. Right: Age-based bias scores across different age groups. Error bars represent standard deviation.

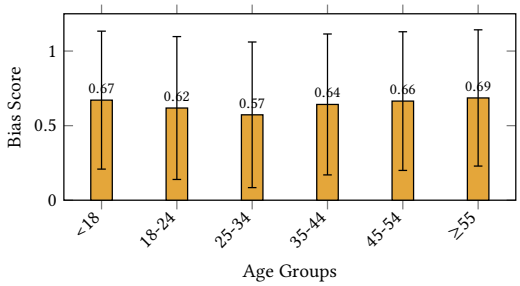

Figure 5: Average bias scores across age demographic groups – Bookcrossing. Error bars represent standard deviation.

The error bars in Figures 3–5 represent standard deviations, providing a measure of variance in bias scores across different items. Wide error bars, particularly visible in age-based bias measurements, suggest that bias effects may be item-dependent - an important consideration for future mitigation strategies.

3.4 Limitations & Broader Implications

While our analysis provides strong evidence about UARS's bias properties, some limitations should be noted. First, our datasets come from specific domains (movies and books), and results may not generalize to all rating contexts. Second, the binary sex classification in available datasets does not reflect the full spectrum of gender identity. Finally, age group boundaries are selected according to the usual groups and different groupings might reveal different patterns.

Moreover, this study highlights key challenges for system designers, regulators, and end-users in balancing user-agnostic principles with fairness considerations. While user-agnostic systems reduce profiling risks, our findings show they may inadvertently amplify disparities across demographic groups, especially age. Designers should carefully evaluate trade-offs between simplicity and fairness, considering demographic-aware adjustments to statistical filtering methods. Regulators may also benefit from integrating domain-specific fairness benchmarks, such as acceptable bias thresholds, to better guide algorithmic evaluations. Finally, transparent explanations of algorithmic decisions are crucial to ensure end-users, particularly those in underrepresented groups, understand and trust ranking systems.

4 CONCLUSIONS AND FUTURE WORK

Our analysis reveals key differences in fairness properties between reputation-based ranking and UARS approaches. Traditional reputation-based systems show systematic biases [14, 16], with disparities in reputation scores consistently disadvantaging certain demographic groups. For sex, reputation scores show significant differences between male and female users, while age-based disparities increase systematically with age difference between groups.

UARS, in contrast, demonstrates more equitable treatment through its statistical filtering mechanism. While not completely eliminating bias, it maintains significantly lower bias scores for sex (0.09-0.16) and shows more balanced treatment of different age groups. The key distinction lies in UARS's user-agnostic design - by focusing solely on rating distributions rather than user attributes, it naturally reduces the opportunity for systematic discrimination.

However, our findings also highlight that complete fairness remains challenging even with statistical filtering. UARS still shows some increased bias for extreme age groups, though at lower levels than reputation-based approaches. This suggests that while statistical filtering provides better protection against discrimination than explicit user profiling, additional considerations may be needed to achieve truly equitable ranking outcomes for all demographic segments.

Looking ahead, research should explore ways to maintain UARS's benefits while addressing these demographic disparities. This could involve developing age-aware statistical filtering criteria that account for group-specific rating patterns, while investigating the underlying causes of different rating behaviors across age groups.

REFERENCES

[1] Judith A Chevalier and Dina Mayzlin. 2006. The effect of word of mouth on sales: Online book reviews. *Journal of marketing research* 43, 3 (2006), 345–354.
[2] Alanah Davis and Deepak Khazanchi. 2008. An empirical study of online word of mouth as a predictor for multi-product category e-commerce sales. *Electronic markets* 18, 2 (2008), 130–141.
[3] Cristobald De Kerchove and Paul Van Dooren. 2010. Iterative filtering in reputation systems. *SIAM J. Matrix Anal. Appl.* 31, 4 (2010), 1812–1834.

[4] Peter De Maeyer. 2012. Impact of online consumer reviews on sales and price strategies: A review and directions for future research. *Journal of Product & Brand Management* 21, 2 (2012), 132–139.

[5] Umberto Grandi and Paolo Turrini. 2016. A network-based rating system and its resistance to bribery. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*. International Joint Conferences on Artificial Intelligence Organization, California, USA, 301–307.

[6] F Maxwell Harper and Joseph A Konstan. 2015. The movielens datasets: History and context. *Acm transactions on interactive intelligent systems (tiis)* 5, 4 (2015), 1–19.

[7] Nan Hu, Indranil Bose, Noi Sian Koh, and Ling Liu. 2012. Manipulation of online reviews: An analysis of ratings, readability, and sentiments. *Decision support systems* 52, 3 (2012), 674–684.

[8] Nan Hu, Paul A Pavlou, and Jennifer Zhang. 2006. Can online reviews reveal a product's true quality? Empirical findings and analytical modeling of online word-of-mouth communication. In *Proceedings of the 7th ACM conference on Electronic commerce*. Association for Computing Machinery, New York, NY, USA, 324–330.

[9] Jan Kietzmann and Ana Canhoto. 2013. Bittersweet! Understanding and managing electronic word of mouth. *Journal of Public Affairs* 13, 2 (2013), 146–159.

[10] Rong-Hua Li, Jeffery Xu Yu, Xin Huang, and Hong Cheng. 2012. Robust reputation-based ranking on bipartite rating networks. In *Proceedings of the 2012 SIAM international conference on data mining*. SIAM, Society for Industrial and Applied Mathematics (SIAM), Anaheim, California, USA, 612–623.

[11] Ewa Maslowska, Edward C. Malthouse, and Stefan F. Bernritter. 2017. *The Effect of Online Customer Reviews' Characteristics on Sales*. Springer Fachmedien Wiesbaden, Wiesbaden, 87–100. https://doi.org/10.1007/978-3-658-15220-8_8

[12] Larry Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. 1997. *PageRank: Bringing order to the web*. Technical Report. Stanford Digital Libraries Working Paper.

[13] European Parliament. 2023. Artificial Intelligence Act. https://www.europarl.europa.eu/doceo/document/TA-9-2023-0236_EN.pdf. Accessed: 2024-02-06.

[14] Guilherme Ramos and Ludovico Boratto. 2020. Reputation (in) dependence in ranking systems: Demographics influence over output disparities. In *Proceedings of the 43rd international ACM SIGIR conference on Research and Development in Information Retrieval*. Association for Computing Machinery, New York, NY, USA, 2061–2064.

[15] Guilherme Ramos, Ludovico Boratto, and Carlos Caleiro. 2020. On the negative impact of social influence in recommender systems: A study of bribery in collaborative hybrid algorithms. *Information Processing & Management* 57, 2 (2020), 102058.

[16] Guilherme Ramos, Ludovico Boratto, and Mirko Marras. 2021. Reputation equity in ranking systems. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*. Association for Computing Machinery, New York, NY, USA, 3378–3382.

[17] Guilherme Ramos, Ludovico Boratto, and Mirko Marras. 2022. Robust reputation independence in ranking systems for multiple sensitive attributes. *Machine Learning* 111, 10 (2022), 3769–3796.

[18] Guilherme Ramos, Mirko Marras, and Ludovico Boratto. 2024. Towards ethical item ranking: A paradigm shift from user-centric to item-centric approaches. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. Association for Computing Machinery, New York, NY, USA, 2667–2671.

[19] João Saúde, Guilherme Ramos, Ludovico Boratto, and Carlos Caleiro. 2021. A robust reputation-based group ranking system and its resistance to bribery. *ACM Transactions on Knowledge Discovery from Data (TKDD)* 16, 2 (2021), 1–35.

[20] João Saúde, Guilherme Ramos, Carlos Caleiro, and Soummya Kar. 2017. Reputation-based ranking systems and their resistance to bribery. In *2017 IEEE International Conference on Data Mining (ICDM)*. IEEE, Institute of Electrical and Electronics Engineers (IEEE), New Orleans, Luisiana, USA, 1063–1068.

[21] Markus Schedl, Emilia Gómez, and Elisabeth Lex. 2023. Trustworthy algorithmic ranking systems. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*. Association for Computing Machinery, New York, NY, USA, 1240–1243.

[22] Sandra Wachter. 2018. Normative challenges of identification in the Internet of Things: Privacy, profiling, discrimination, and the GDPR. *Computer law & security review* 34, 3 (2018), 436–449.

[23] Shoujin Wang, Xiuzhen Zhang, Yan Wang, and Francesco Ricci. 2024. Trustworthy recommender systems. *ACM Transactions on Intelligent Systems and Technology* 15, 4 (2024), 1–20.

[24] Cai-Nicolas Ziegler, Sean M McNee, Joseph A Konstan, and Georg Lausen. 2005. Improving recommendation lists through topic diversification. In *Proceedings of the 14th international conference on World Wide Web*. Association for Computing Machinery, New York, NY, USA, 22–32.